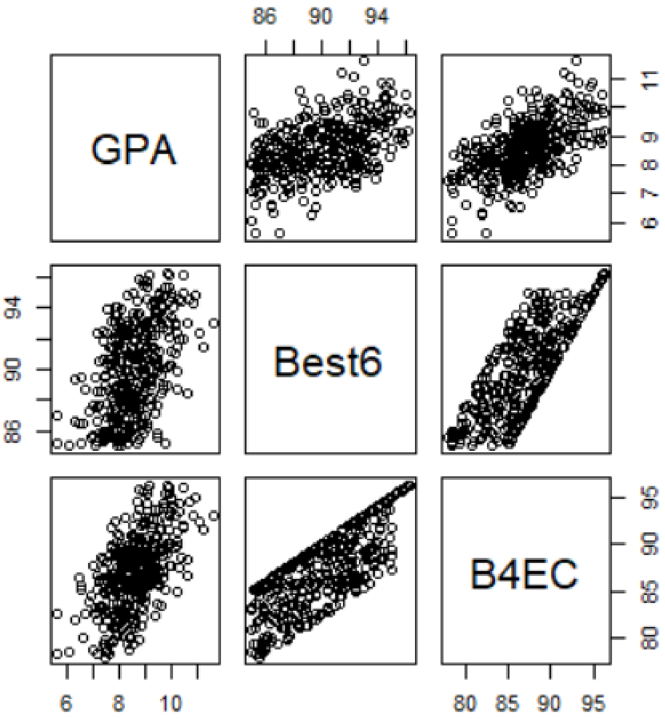
工管三 B08701119 張廷鋒

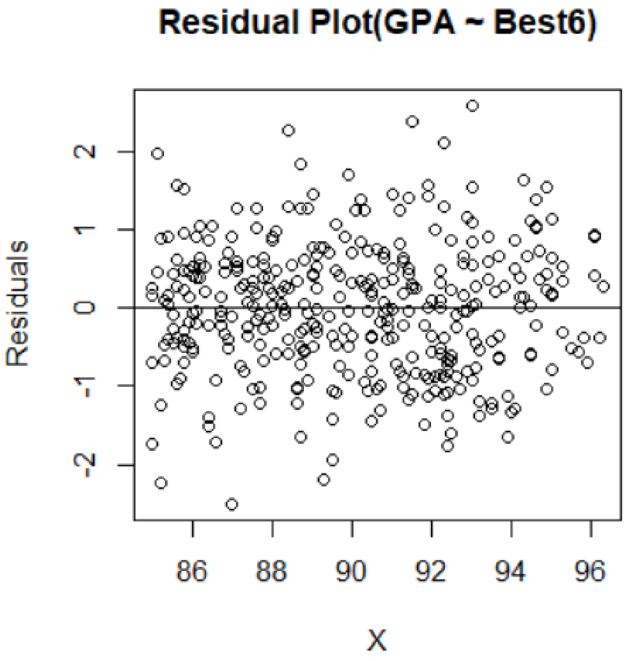
1. Ontario high school students must complete a minimum of six Ontario Academic Credits (OACs) to gain admission to a university in the province. Most students take more than six OACs because universities take the advantage of the best six in deciding which students to admit. Most programs at universities require high school students to select certain courses. For example, science programs require two of chemistry, biology, and physics. Students applying to engineering must complete at least two mathematics OACs as well as physics. In recent years, one business program began an examination of all aspects of its program, including the criteria used to admit students. Students are required to take English and calculus OACs, and the minimum high school average is about 85%. Strangely enough, even though students are required to complete English and calculus, the marks in these subjects are not included in the average unless they are in the top six courses in a student’s transcript. To examine the issue, the registrar took a random sample of students who recently graduated with the BBA (bachelor of business administration degree). He recorded the university GPA (range 0 to 12), the high school average based on the best sic courses, and the high school average using English and calculus and the four next best marks in the data file “OCAs.txt”.

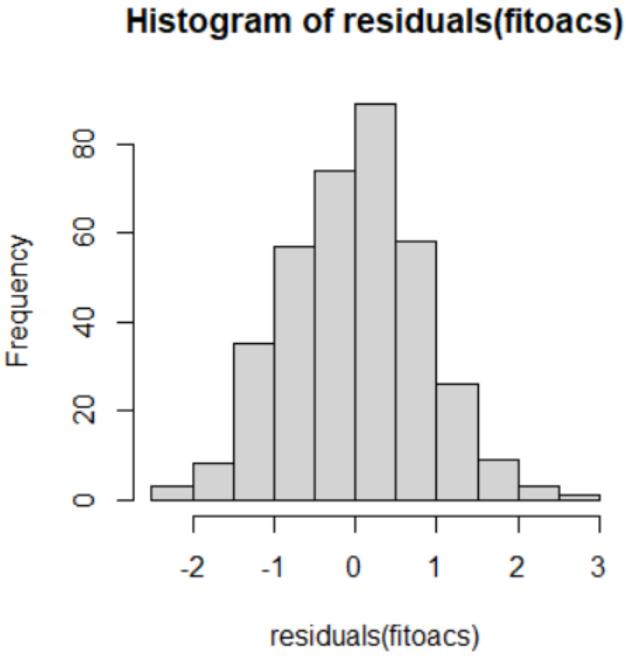


Pairplot 中可以看出 GPA 和兩變數都各具有正向相關。

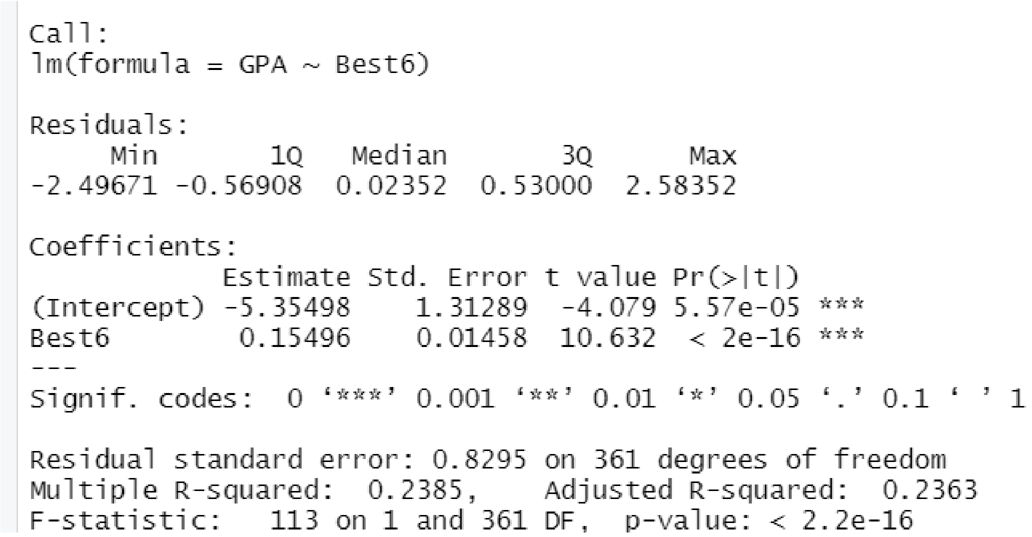
(a) Is there a relationship between university grades and high school average using the best six OACs?

在檢視模型 summary 之前，先使用 residual plot 和 histogram 檢查是否符合 3+1 個假設。





Residual plot 顯示殘差符合平均為 0、變異數為常數及彼此獨立之假設，而 Histogram 顯示殘差符合常態分佈，3+1 個假設都符合。



```
Call:
lm(formula = GPA ~ Best6)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.49671 -0.56908  0.02352  0.53000  2.58352 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -5.35498    1.31289  -4.079 5.57e-05 ***
Best6        0.15496    0.01458  10.632  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8295 on 361 degrees of freedom
Multiple R-squared:  0.2385,	Adjusted R-squared:  0.2363 
F-statistic:   113 on 1 and 361 DF,  p-value: < 2.2e-16
```

以下假設所有 test 的 significance level = 0.05。

首先先觀察 F-test 的部分，p-value 為 2.2e-16，顯著小於 0.05，因此可以拒絕 H0: B0=B1=0 的虛無假設，推斷此模型係數不全為 0 具有解釋力。

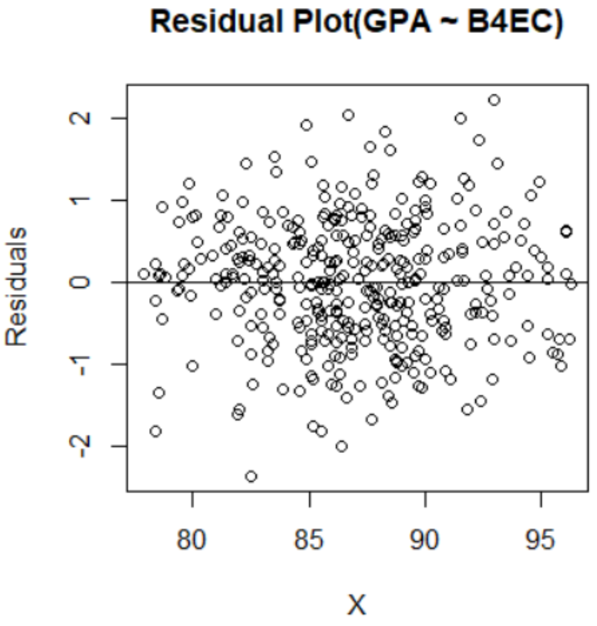
接著可以看到 COEFFICIENTS 的部分，B1 的 p-value=2e-16，顯著小於 0.05，因此可拒絕 B1=0 的虛無假設，可見此模型唯一的係數 B1 具有解釋力。

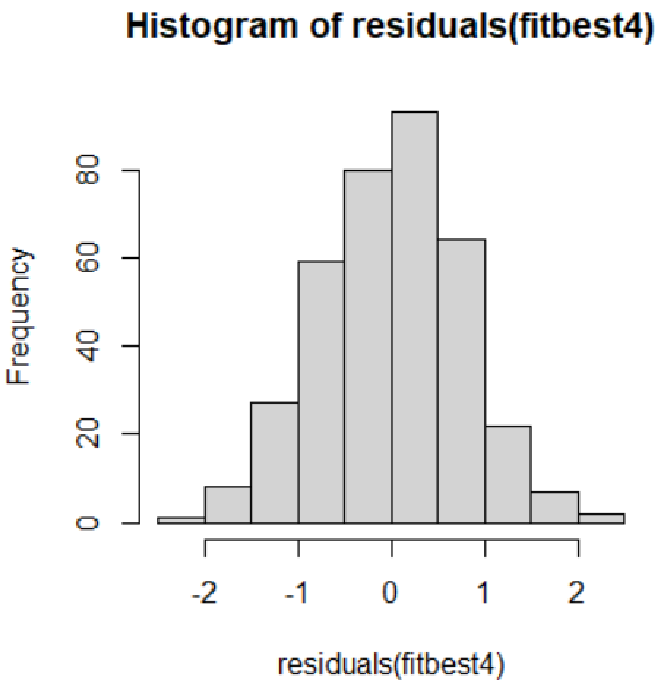
最後看到 Multiple R-squared 係數為 0.2385，代表 y 方向的變異有 23.85%由 X 方向變異所解釋，開根號得出相關係數 r = 0.4884，具有一定的正相關性。

回歸方程式為：y = -5.35498 + 0.15496 * Best6

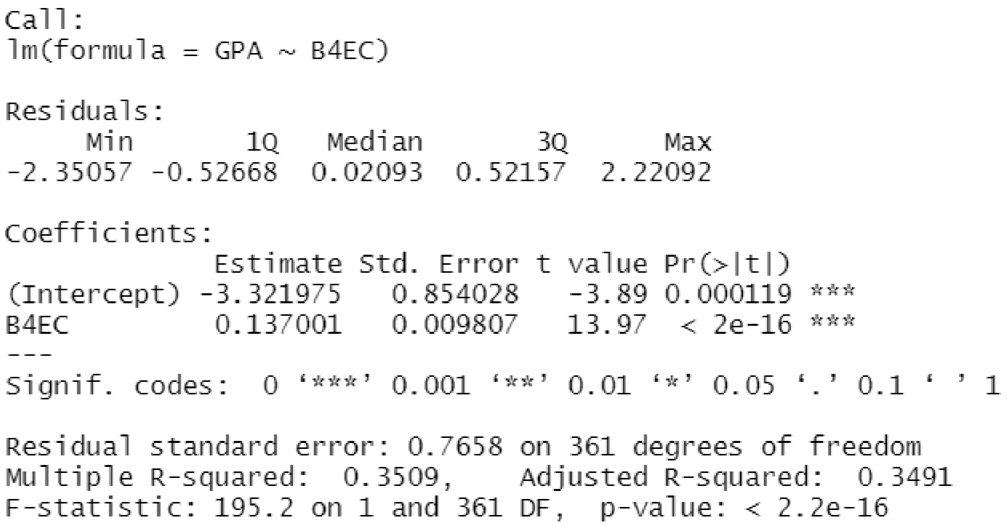
(b) Is there a relationship between university grades and high school average using the best four OACs?

在檢視模型 summary 之前，先使用 residual plot 和 histogram 檢查是否符合 3+1 個假設。





Residual plot 顯示殘差符合平均為 0、變異數為常數及彼此獨立之假設，而 Histogram 顯示殘差符合常態分佈，3+1 個假設都符合。



```
Call:
lm(formula = GPA ~ B4EC)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.35057 -0.52668  0.02093  0.52157  2.22092 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -3.321975   0.854028   -3.89 0.000119 ***
B4EC         0.137001   0.009807   13.97  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7658 on 361 degrees of freedom
Multiple R-squared:  0.3509,    Adjusted R-squared:  0.3491 
F-statistic: 195.2 on 1 and 361 DF,  p-value: < 2.2e-16
```

首先先觀察 F-test 的部分，p-value 為 2.2e-16，顯著小於 0.05，因此可以拒絕 H0: B0=B1=0 的虛無假設，推斷此模型係數不全為 0 具有解釋力。

接著可以看到 COEFFICIENTS 的部分，B1 的 p-value=2e-16，顯著小於 0.05，因此可拒絕 B1=0 的虛無假設，可見此模型唯一的係數 B1 具有解釋力。

最後看到 Multiple R-squared 係數為 0.3509，代表 y 方向的變異有 35.09%由 X 方向變異所解釋，開根號得出相關係數 r = 0.5924，具有一定的正相關性。

回歸方程式為：y = -3.321975 + 0.137001 * B4EC

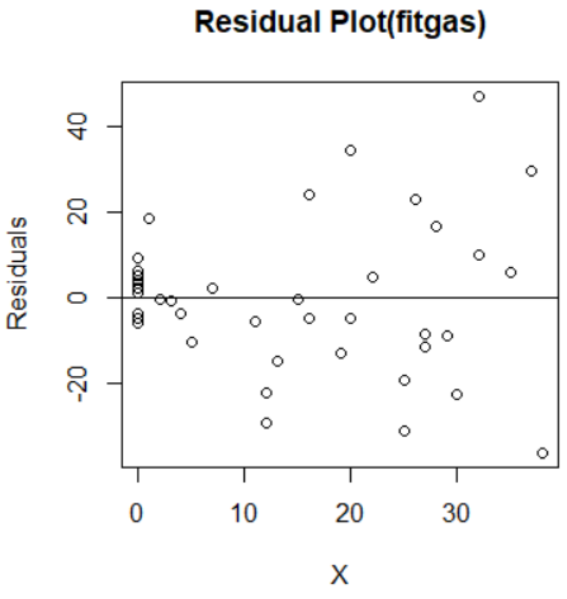
(c) Write a comment to the university's academic vice president describing your statistical analysis and your recommendations.

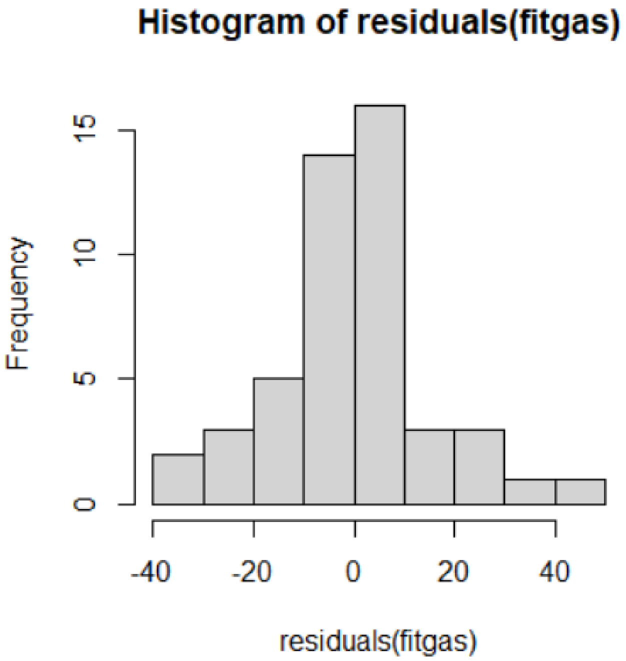
Best6 和 B4EC 這兩個模型及係數都具備顯著的解釋力也都符合 3+1 個假設，是可用的模型。不過 B4EC 這個模型呈現出較高的 R2，可能顯示其對於 y（大學成績）具備較好的解釋力。由於樣本數偏少，故我會建議在蒐集更多 data 再跑模型結果會更貼近現實。若無法得到更多 data，則我會建議採用 B4+E+C 為審核標準，因為從現有的數據中觀察，這個項目與大學成績的相關性較高。

2. Utility companies in many older communities still rely on "meter readers" who visit homes to read meters that measure consumption of electricity and gas. Unless someone is home to let the meter reader inside, the utility company has to estimate the amount of energy used. The utility company in this example sells natural gas to homes in the Philadelphia area. Many of these are older homes that have the gas meter in the basement. We can estimate the use of gas in these homes with a simple linear model. The explanatory variable is the average number of degrees below 65 during the billing period, and the response is the number of hundred cubic feet of natural gas (CCF) consumed during the billing period (about a month). The explanatory variable is set to 0 if the average temperature is above 65 (assuming a homeowner won't need heating in this case).

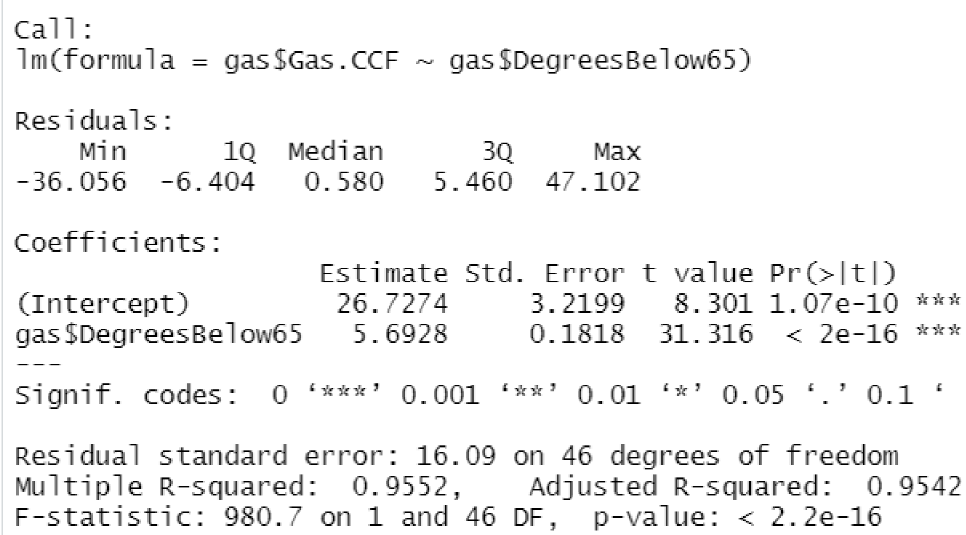
(a) Use R to fit a simple linear model with the date "gas_consumption.txt". Do the analysis, make the plot, and summarize the results.

在檢視模型 summary 之前，先使用 residual plot 和 histogram 檢查是否符合 3+1 個假設。





Residual plot 顯示殘差符合平均為 0、變異數為常數及彼此獨立之假設，而 Histogram 顯示殘差符合常態分佈，3+1 個假設都符合。

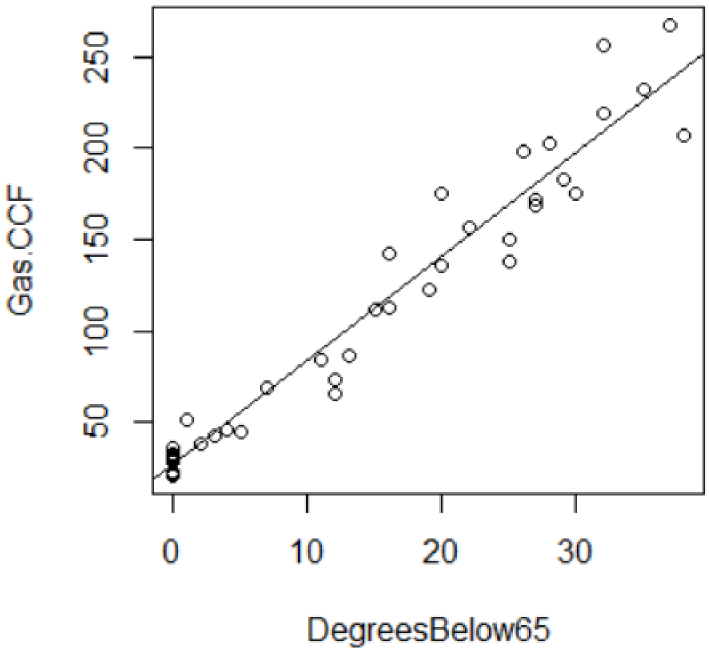


```
Call:
lm(formula = gas$Gas.CCF ~ gas$DegreesBelow65)

Residuals:
    Min      1Q  Median      3Q     Max 
-36.056  -6.404   0.580   5.460  47.102 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)         26.7274     3.2199   8.301 1.07e-10 ***
gas$DegreesBelow65   5.6928     0.1818  31.316  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.09 on 46 degrees of freedom
Multiple R-squared:  0.9552,	Adjusted R-squared:  0.9542 
F-statistic: 980.7 on 1 and 46 DF,  p-value: < 2.2e-16
```

首先先觀察 F-test 的部分，p-value 為 2.2e-16，顯著小於 0.05，因此可以拒絕 H0: B0=B1=0 的虛無假設，推斷此模型係數不全為 0 具有解釋力。

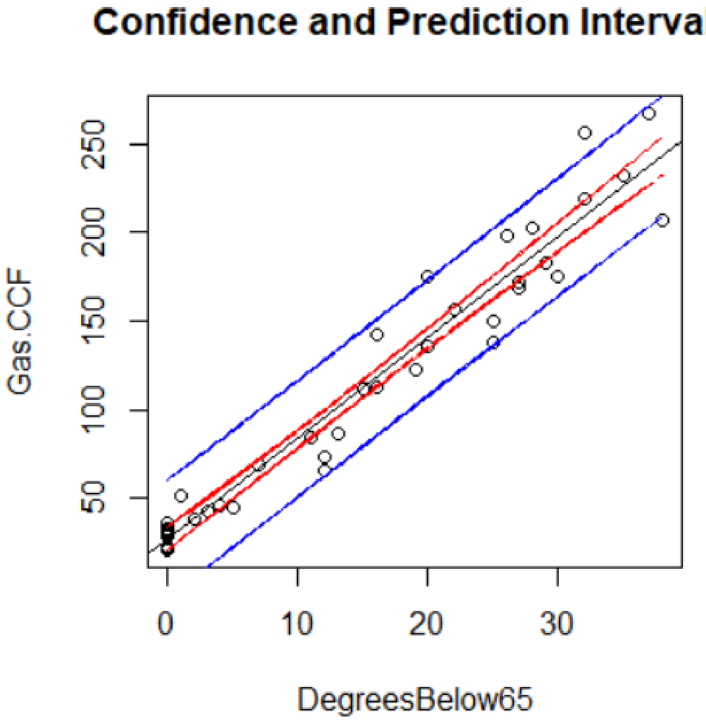
接著可以看到 COEFFICIENTS 的部分，B1 的 p-value=2e-16，顯著小於 0.05，因此可拒絕 B1=0 的虛無假設，可見此模型唯一的係數 B1 具有解釋力。

最後看到 Multiple R-squared 係數為 0.9552，代表 y 方向的變異有 95.52%由 X 方向變異所解釋，開根號得出相關係數 r = 0.9773，具有非常高的正相關性。

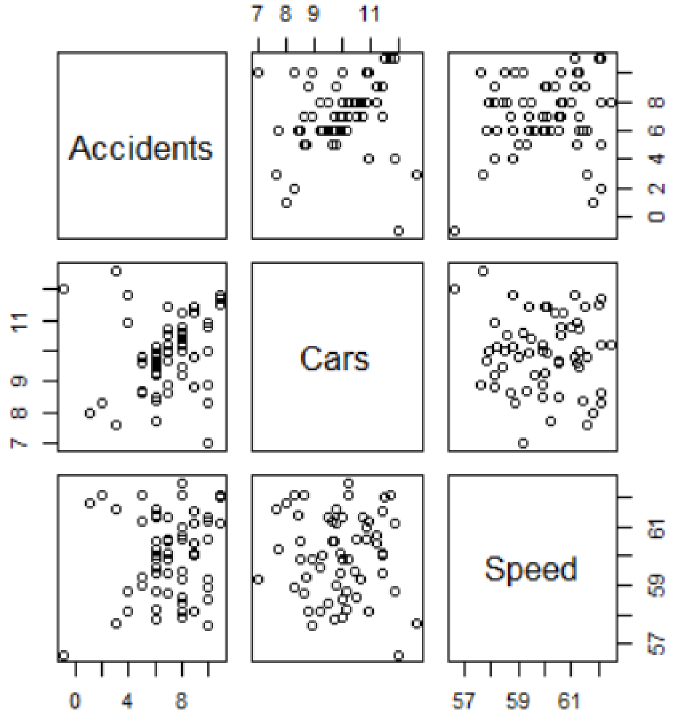


回歸方程式為：y = 26.7274 + 5.6928 * B4EC，從這張散步圖中也可看出多數的點皆聚集在回歸線附近，可見應變數和自變數之間相關性極高。

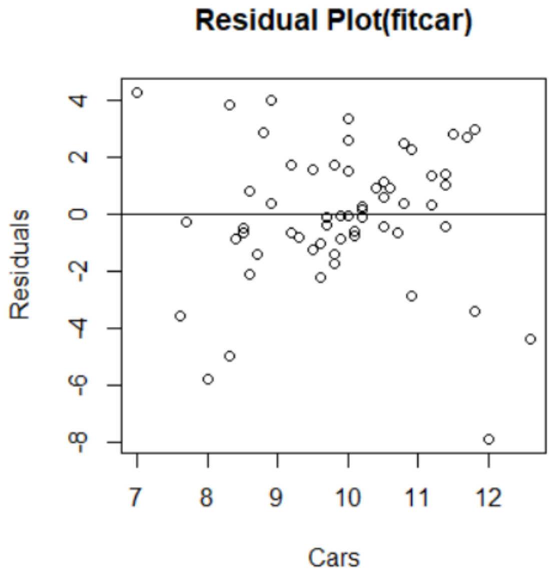
(b) Modify the script provided for this lecture to create the “Confidence and Prediction Intervals” plot shown on lecture note p. 2-32 and re-printed below.

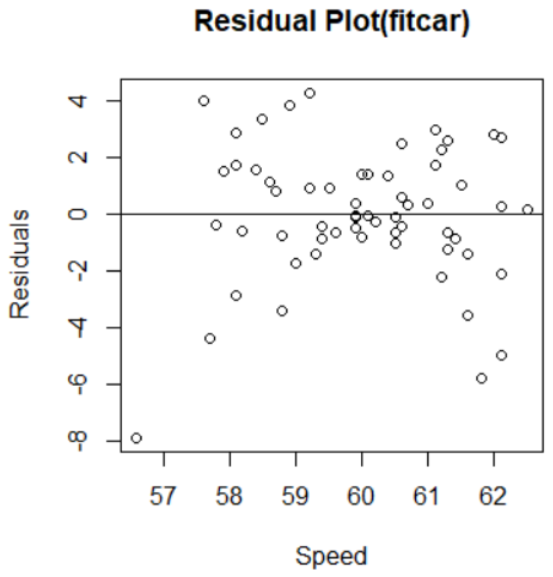


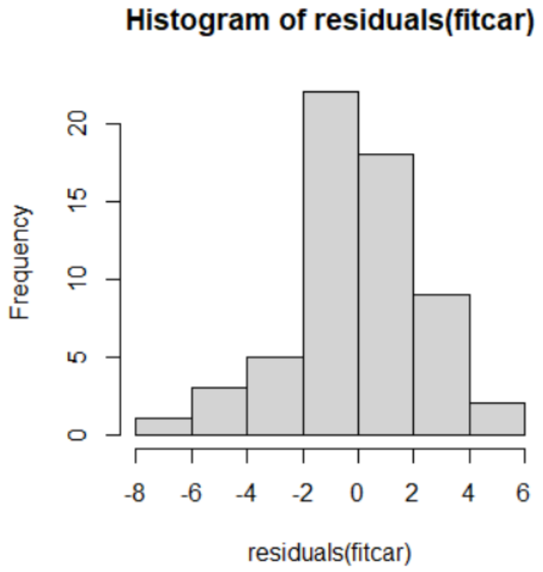
3. The number of car accidents on a particular stretch of highway seems to be related to the number of vehicles that travel over it and the speed at which they are traveling. A city alderman decided to ask the county sheriff to provide him with statistics covering the last few years, with the intention of examining these data statistically so that he can (if possible) introduce new speed laws that will reduce traffic accidents. Using the number of accidents as the dependent variable, he obtained estimates of the number of cars passing along a stretch of road and their average speeds (in miles per hour). The observations for 60 randomly selected days were recorded in file “car_accident.txt”. Find a linear model that appropriately reveals the relationship among these variables. Clearly state your model building process and interpret the results of your final model.



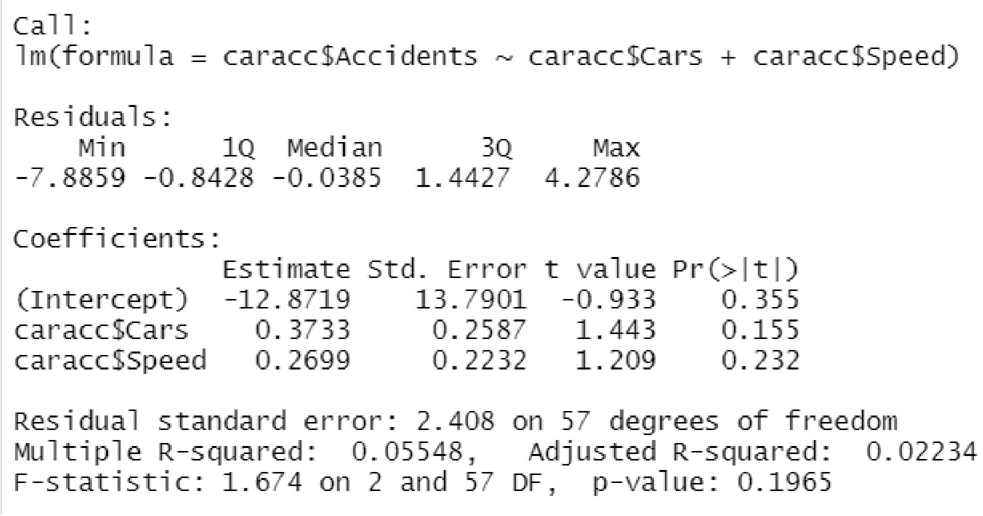
從 pairplot 中可以看出，cars 和 speed 並沒有明顯的相關性。
接下來先用這兩個變數建立 multiple LR 模型。


Residual Plot(fitcar)
Residuals
4
2
0
-2
-4
-6
-8
7
8
9
10
11
12
Cars


Residual Plot(fitcar)
Residuals
4
2
0
-2
-4
-6
-8
57
58
59
60
61
62
Speed


Histogram of residuals(fitcar)
Frequency
20
15
10
5
0
-8
-6
-4
-2
0
2
4
6
residuals(fitcar)

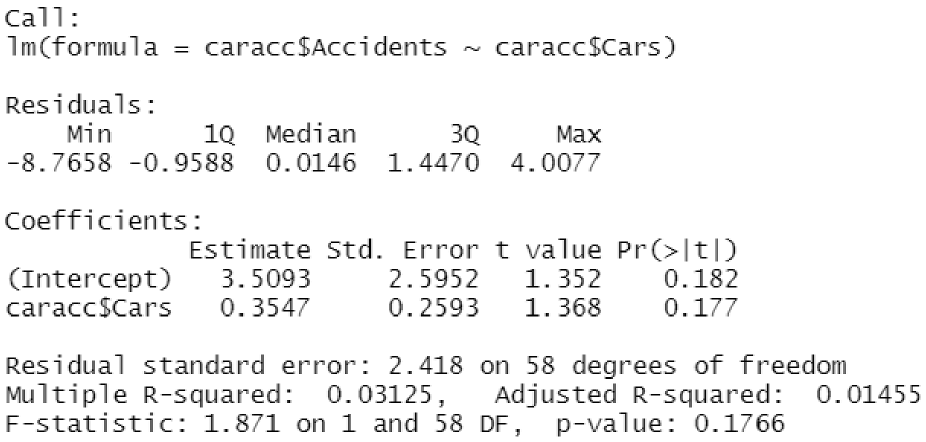
Residual plot 顯示殘差符合平均為 0、變異數為常數及彼此獨立之假設，而 Histogram 顯示殘差符合常態分佈，3+1 個假設都符合。



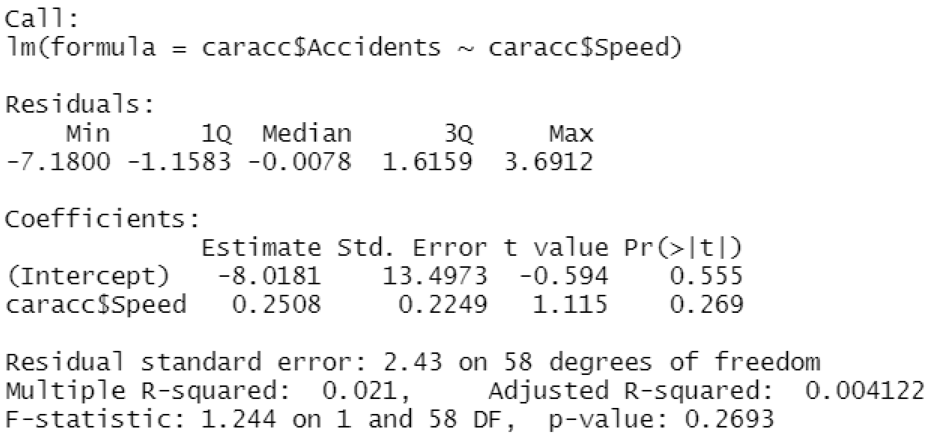
```
Call:
lm(formula = caracc$Accidents ~ caracc$Cars + caracc$Speed)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.8859 -0.8428 -0.0385  1.4427  4.2786 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -12.8719    13.7901  -0.933    0.355
caracc$Cars    0.3733     0.2587   1.443    0.155
caracc$Speed   0.2699     0.2232   1.209    0.232

Residual standard error: 2.408 on 57 degrees of freedom
Multiple R-squared:  0.05548,	Adjusted R-squared:  0.02234 
F-statistic: 1.674 on 2 and 57 DF,  p-value: 0.1965
```

首先先觀察 Global Usefulness test 的部分，p-value 為 0.1965，大於 0.05，因此可以接受 H0: B0=B1=0 的虛無假設，推斷此模型係數全為 0 不具有解釋力。

將自變數單獨建立模型觀察效果如何。



```
Call:
lm(formula = caracc$Accidents ~ caracc$Cars)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.7658 -0.9588  0.0146  1.4470  4.0077 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.5093     2.5952   1.352    0.182
caracc$Cars   0.3547     0.2593   1.368    0.177

Residual standard error: 2.418 on 58 degrees of freedom
Multiple R-squared:  0.03125,	Adjusted R-squared:  0.01455 
F-statistic: 1.871 on 1 and 58 DF,  p-value: 0.1766
```



```
Call:
lm(formula = caracc$Accidents ~ caracc$Speed)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.1800 -1.1583 -0.0078  1.6159  3.6912 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -8.0181    13.4973  -0.594    0.555
caracc$Speed   0.2508     0.2249   1.115    0.269

Residual standard error: 2.43 on 58 degrees of freedom
Multiple R-squared:  0.021,	Adjusted R-squared:  0.004122 
F-statistic: 1.244 on 1 and 58 DF,  p-value: 0.2693
```

三個模型皆沒有通過 Global Usefulness test。比較各模型的 Adjusted R2，可見第一個模型仍表現得最好。儘管他並沒有通過 Global Usefulness test，還是會選擇第一個模型。